Article

American Journal of Evaluation
32(1) 8-28

DOI: 10.1177/1098214010378354
http://aje.sagepub.com

Measures of Program Effectiveness Based on Retrospective Pretest Data: Are All Created Equal?

Kim Nimon[1], Drea Zigarmi[2], and Jeff Allen[1]

Abstract

This study analyzed data from four evaluation designs incorporating the retrospective pretest (i.e., thentest), analyzing the effects of self-report pretesting and post-program survey format on a set of self-report measures. Validity of self-report data was assessed by comparing the criterion validity of then ratings to established benchmarks and by including a control condition. This study found that designs incorporating separate posttest and thentest surveys yielded the most comparable levels of criterion validity for then ratings and the least biased measures of program effectiveness. Conversely, designs that incorporated a single post-program survey, with adjacent post and then items, yielded the least comparable levels of criterion validity for then ratings and the most biased measures of program effectiveness. This study also found a slight interaction between the effects of self-report pretesting and post-program survey format. Implications for program evaluation are discussed.

Keywords

retrospective pretest, experience limitation, pretest sensitization, implicit theory of change

Over the last five decades, researchers have proposed the retrospective pretest as a technique to control for key factors jeopardizing the validity of self-report effects derived from posttest-only control group, cross-sectional, and pre–post designs and to rule out rival hypotheses. Citing studies conducted by Deutsch and Collins (1951), Sears, Maccoby, and Levin (1957), and Walk (1956), Campbell and Stanley (1963) recommended supplementing posttest-only control group designs with a retrospective pretest to validate pre-program equivalence of experimental and control groups and rule out the rival hypothesis of selection bias between groups. They further identified the use of retrospective pretests to partially curb rival hypotheses of history, selective mortality, and shifts in initial selection that might confound effects detected in cross-sectional designs comparing different age groups (e.g., freshman and seniors) at the same time. Almost two decades later, Howard et al.

[1]University of North Texas, Dallas, TX, USA
[2]Ken Blanchard Company, Escondido, CA, USA

Corresponding Author:
Kim Nimon, 18352 Dallas Parkway, # 136-407, Dallas, TX 75287, USA
Email: kim.nimon@gmail.com

(1979) proposed extending pre-post designs by *adding* the retrospective pretest to identify and mitigate the effect of response–shift bias when evaluating program effectiveness. A little over a decade later, Aiken and West (1990) echoed the propositions set forth by Howard et al. (1979) and extended their work by demonstrating how the retrospective pretest could be used as a tool to detect sources of bias in self-report pretest measures, including experience limitation, condition justification, altered states, and self-presentation. Within the last decade, evaluators (e.g., Pelfrey & Pelfrey, 2009) have advocated *replacing* the traditional pretest in pre-post designs with the retrospective pretest as a practical and valid means to determine program outcomes, controlling for the effects of response–shift bias and sensitization effects.

In program evaluation, the retrospective pretest distinguishes itself from the traditional pretest via its relationship to the program. Also called the thentest since Howard (1980), the retrospective pretest is administered after the program and tasks participants to respond to survey items as they were "then" (i.e., before the program). Incorporating the thentest to evaluate pre- and post-program differences solicits a validity conundrum (Hill & Betz, 2005). On one hand, the thentest may provide more valid results than a traditional pretest if, at the time of the traditional pretest, participants (a) lack familiarity with the dimension of self-rating (i.e., experience limitation), (b) unconsciously exaggerate self-ratings to justify their emotional state (i.e., condition justification), (c) are in a medical state (e.g., drug induced) that prevents accurate self-ratings (i.e., altered state), or (d) consciously distort self-ratings to access desired training (i.e., self-presentation) (Aiken & West, 1990). On the other hand, thentest data may be biased, if participants reconstruct their memories as a function of (a) exaggerating the program's effect to justify the effort expended in the program (i.e., effort justification), (b) assuming that the program had the desired effect (i.e., implicit theory of change), or (c) enhancing the degree to which they have improved as a means of impression management (i.e., self enhancement) (Taylor, Russ-Eft, & Taylor, 2009). Participants' recall may also bias thentest data, if length and specificity of the time period diminishes the recall process (Pratt, McGuigan, & Katzev, 2000).

Beyond the traditional validity concerns of thentest data, the various formats and encompassing designs threaten the validity of the data collected post-program. Formatting post-program self-report assessment with adjacent post and then items on a single survey may elicit biased post and then ratings. In addition, any derived effects (e.g., program and response–shift) are potential targets of bias. Moreover, if a traditional pre-post design incorporates a thentest, self-report pretesting presents a potential source of bias in post-program self-assessment data. Although administering a traditional pretest in addition to the thentest facilitates the detection of response–shift (i.e., differences between then and pre ratings) (Howard, 1980; Pratt et al., 2000), the detection of response –shift may come at the price of carryover effects in post ratings, interference effects in post and then ratings, or sensitization to how participants respond to the program and post-program assessments. In addition, any effects derived from biased post and then ratings are potential targets of self-report pretest effects. The following sections present theoretical and empirical support for post-program survey format and self-report pretesting effects on post and then ratings. We also present findings from literature that demonstrate a possible interaction between post-program survey format and self-report pretesting.

Post-Program Survey Format

Schwarz (1996) provides theoretical support for a single post-program survey, with adjacent post and then items, introducing bias in resulting post and then ratings. Situating survey design within Grice's (1975) maxims of conversation, Schwarz argued that subjects use contextual information in interpreting survey items, relating the item to the context of an ongoing exchange. More specifically, Schwarz suggested that subjects consider the content of adjacent items in the process of interpreting a question's intended meaning. Schwarz (1999) provided empirical data to support his theory when he found that whether subjects concluded that marital satisfaction was a major or a minor

contributor to general life satisfaction depended on the order of the questions, where correlations ranged from .18 to .67 as a function of question order.

Evaluators following post-program survey procedures recommended by Howard et al. (1979) collect post and then ratings via the same questionnaire and task participants to respond to then items relative to their post responses. Howard et al. described the post-program survey procedure as follows:

> First, they were to report how they perceived themselves to be at present (Post). Immediately after answering each item in this manner, they were to answer the same item again, this time in reference to how they now perceived themselves to have been just before the workshop was conducted (Then). Subjects were instructed to make the Then response in relation to the corresponding Post response to insure that both responses would be made from the same perspective. (p. 5)

Such a post-program survey procedure creates a common referent for how participants respond to post and then items. However, placing post and then items side by side and tasking participants to respond to then items relative to their post responses creates a contextual effect in which participants attend to the contrast between the two ratings. This contextual effect could in turn affect how participants reconstruct their initial status and bias post-program measurement outcomes. For example, personal recall theory suggests that if individuals are expecting to experience a change in knowledge, skills, or attitudes (KSAs) as a result of attending a training program, individuals may reconstruct their initial status in concert with an implicit theory of change and indicate a program effect (i.e., a practical difference between post and then ratings), even when no such change occurs (Ross, 1989). Similarly, impression management theory suggests that if individuals believe that the appearance of improvement will make them look good and please the program leader, individuals may moderate their initial status to generate an impression of improvement, even if no improvement occurs (Pearson, Ross, & Dawes, 1992). As well, cognitive dissonance theory suggests that if individuals perceive no positive effects from participating in an intervention, they may reconstruct their initial status to avoid the cognitive dissonance associated with the time and effort they invested in the program (Hill & Betz, 2005). Although such inflationary biases are possible when participants respond to separate posttests and thentests, the post-program survey procedure outlined by Howard et al. (1979) may facilitate such biases if participants attend to their post ratings as they reconstruct their then ratings.

Across peer-reviewed journal articles reporting on retrospective pretest designs, one study (Terborg & Davis, 1982) was found to have considered the impact of post-program survey format. Terborg and Davis analyzed then ratings and found no statistical difference when administering the post-program assessment with one or two surveys. However, analysis of the reported means indicates that then ratings tended to be higher for the post-program assessment with one survey.

Self-Report Pretesting

Threats to validity associated with self-report pretesting in traditional pre-post designs provide theoretical support for self-report pretesting introducing bias into post-program data from pre-post designs incorporating a thentest. Lam and Bengo (2003) identified three potential sources of bias in traditional pre-post designs attributable to self-report pretest effects in post-program data—carryover, interference, and sensitization. They associated a carryover effect with a positive bias in post ratings as a consequence of participants attempting to correct for errors in their pretest responses. They associated an interference effect with a negative bias in post ratings as a consequence of repetitive testing. They associated a sensitization effect with a bias in how participants respond to the intervention and the posttest as a consequence of the pretest making certain aspects of the program more salient.

By extending the scholarship of Lam and Bengo (2003) on the effects of self-report pretesting in traditional pre-post designs to designs in which a thentest is also incorporated, one may hypothesize self-report pretest effects in post and then ratings. Self-report pretesting may introduce effects in post-program data from retrospective pretest designs similar to those hypothesized in traditional pre-post designs. These effects include: (a) a carryover effect in post ratings, if a negative response–shift occurs in how participants perceive their initial status and participants are unaware that they will have an opportunity to reevaluate their pre-program KSAs, (b) an interference effect in post as well as then ratings, if the repetitive testing causes participants to feel bored or fatigued, and (c) a sensitization effect in post and/or then ratings, if the pretesting makes certain aspects of the program more salient (Sprangers & Hoogstraten, 1989; Terborg & Davis, 1982).

Although empirical studies examining self-report pretest effects in post ratings have been published since as far back as the 1950s (Lam & Bengo, 2003), published studies examining self-report pretest effects in retrospective pretest designs are limited. Sprangers and Hoogstraten (1989) found that self-report pretesting positively biased post and then ratings. Sprangers and Hoogstraten reasoned that this effect was due to selective attention as opposed to compliance with implicit demands, given that the resulting gains from post and then ratings were *lower* than gain scores from participants who did not take a traditional pretest.

Terborg and Davis (1982) conducted a similar experiment and reported no self-report pretest effect on post or then ratings. However, analysis of the reported means indicated a similar trend found by Sprangers and Hoogstraten (1989). Both post and then ratings were higher (although not statistically significant) for participants who completed a self-report pretest. However, in contrast to what Sprangers and Hoogstraten found, the resulting gains from post and then ratings were *higher* than the gain scores of participants who did not take a self-report pretest.

Findings from Sprangers and Hoogstraten (1989) and Terborg and Davis (1982) provide inconclusive data regarding the effect of self-report pretesting on gains between post and then ratings. However, the disparate findings may be a function of post-program survey administration. Terborg and Davis (1982, p. 119) collected post and then ratings "simultaneously on the same questionnaire in accordance with procedures recommended by Howard et al. (1979)." Sprangers and Hoogstraten contrasted their survey administration to the recommendations set forth in Howard et al. (1979) as follows:

> In line with our earlier research, and in contrast to the procedure of Howard and his colleagues, subjects first completed all posttest items conventionally and, while keeping the posttest in front of them, reported how they now perceived themselves to have been prior to the training. The instruction stated, just as in Howard's work, that subjects were to answer each retrospective item in relation to the corresponding posttest item, starting with Item 1. (p. 149)

Post-Program Survey Format and Self-Report Pretesting Interaction

Findings from Sprangers and Hoogstraten (1989) and Terborg and Davis (1982) suggest the possibility of an interaction between self-report pretesting and post-program survey format. In the case of the "test-after-test" post-program survey procedure (Sprangers & Hoogstraten, 1989, p. 149), the effect of self-report pretesting resulted in lower gains between post and then ratings. In the case of the "simultaneous" collection of post and then responses (Terborg & Davis, 1982, p. 119), the effect of self-report pretesting resulted in higher gains between post and then ratings.

Data of Sprangers and Hoogstraten (1989) suggest that when evaluators instruct participants to respond to all post items on one form before responding to then items on a second form, program effects may be *attenuated*, if participants are predisposed to the expected outcomes of the intervention as a consequence of taking a traditional pretest. In contrast, the data of Terborg and Davis (1982)

suggest that when evaluators instruct participants to respond to post and then items simultaneously, program effects may be *amplified*, if participants are predisposed to the expected outcomes of an intervention as a consequence of taking a traditional pretest. When considering the potential bias associated with then ratings (e.g., implicit theory of change, impression management, and effort justification), the data of Sprangers and Hoogstraten suggest that such bias may in part be controlled by designing a program evaluation with a self-report pretest and separate post-program surveys, while the data of Terborg and Davis suggest that such bias may be facilitated by having participants respond to a self-report pretest and a single post-program survey.

Although grounded theory supports the findings of Sprangers and Hoogstraten (1989) and Terborg and Davis (1982), it is also possible that differences in the effect of self-report pretesting on gains between post and then ratings between the two studies may be attributable to sampling error. In Sprangers and Hoogstraten, the number of participants in each group was 9. In Terborg and Davis, the number of participants in the group who took the traditional pretest was 12 and the number of participants in the group who did not take the traditional pretest was 8. Clearly, more work is required to determine whether and how self-report pretesting interacts with post-program survey administration.

Purpose of the Study

Published research considering the impact of self-report pretesting and post-program survey format in data derived from retrospective pretest designs is limited. The prior studies reviewed have in essence made selective comparisons among variations of four designs incorporating the retrospective pretest:

- Pre-Post-Then: Pre-post design incorporating a thentest in which the posttest and thentest are administered as two separate questionnaires, with the posttest administered before the thentest.
- Pre-Post&then: Pre-post design incorporating a thentest with post and then items administered as a single questionnaire, with post items presented first.
- Post-Then: Posttest design incorporating a thentest in which the posttest and thentest are administered as two separate questionnaires, with the posttest administered before the thentest.
- Post&then: Posttest design incorporating a thentest with post and then items administered as a single questionnaire, with post items presented first.

It appears that no study has been published considering the combined effects of self-report pretesting and post-program survey format and examined measurement outcomes across all four associated designs. In addition, published research on self-report pretesting and post-program survey format has focused primarily on between-group differences in select post-program ratings and has not evaluated within-group measures of program effectiveness or the criterion validity of post-program data.

In retrospective pretest designs, evaluators typically assess program effectiveness by constructing a within-group design and examining the statistical and practical significance between post and then ratings (Lamb & Tschillard, 2005). As in traditional pre-post designs, within-group retrospective pretest designs are more powerful than between-group designs because each participant serves as his or her own control (Pratt et al., 2000). However, published literature has not considered how self-report pretesting and post-program survey format relate to such measures of program effectiveness. Furthermore, the rationale for using then ratings over pre ratings is that then ratings are more valid than pre ratings, given that participants base then ratings on the same standard of measure that they use when providing post ratings (Pratt et al., 2000). Citations to the work of Howard et al. (1979) usually bolster this analytic strategy, as Howard et al. found differences between post and

then ratings correlated higher with objective measures of change than differences between post and pre ratings. However, Howard et al. advised that researchers view their findings with caution because there are fundamental issues with measuring and correlating change scores (cf., Cronbach & Furby, 1970; Linn & Slinde, 1977). What appears to be missing in the literature is an analysis of the criterion validity of observed post-program measures as a function of retrospective pretest design differences (e.g., self-report pretesting and post-program survey format).

The current study sought to fill gaps in the retrospective pretest design literature by examining self-report post-program measurement outcomes from four retrospective pretest designs (Pre-Post-Then, Pre-Post&then, Post-Then, and Post&then), guided by two research questions:

1. Does the level of criterion validity of then ratings differ by retrospective pretest design (Pre-Post-Then, Pre-Post&then, Post-Then, and Post&then)?
2. Do measures of program effectiveness derived from post and then ratings differ by retrospective pretest design (Pre-Post-Then, Pre-Post&then, Post-Then, and Post&then)?

To place our study's finding within context of prior research, we also considered the following:

1. Are pre ratings in the current study subject to experience limitation (Aiken & West, 1990)?

Based on our analysis of literature (Lam & Bengo, 2003; Schwarz, 1996, 1999; Sprangers & Hoogstraten, 1989; Taylor et al., 2009; Terborg & Davis, 1982; Thorndike, 1949), we deemed it was likely that data from the first two research questions would correlate. We hypothesized that designs incorporating separate posttest and thentest surveys would be more prone to bias than designs that included a single post-program survey. We further hypothesized that self-report pretesting would interact with post-program survey format such that the Pre-Post-Then group would produce less bias and the Pre-Post&Then group would produce more bias. Presuming that such bias attenuates criterion validity of then ratings and inflates measures of program effectiveness, we expected that data resulting from the first two research questions would relate. This study also hypothesized that participants would not have sufficient knowledge to accurately judge their pre-program abilities at the time of taking a traditional pretest, resulting in pre ratings that would not conform to established benchmarks and that would exhibit response–shift bias.

Method

Research Design

The study was based on a repeated measures design with a blocking factor (i.e., split-plot or mixed-model design). The blocking factor was representative of a fully crossed 2×2 factor with two levels for self-report pretesting (yes or no) and two levels for post-program survey format (one or two). We modeled the blocking factor in the split-plot design as one variable that represented the fully crossed combinations of self-report pretesting and post-program survey format. The blocking factor therefore represented the four retrospective pretest designs (Pre-Post-Then, Pre-Post&then, Post-Then, and Post&then).

As depicted in Table 1, participants were assigned to one of the four retrospective pretest designs. Class rosters were used to make participant assignments. In the first class assigned, the first participant listed in the class roster was assigned to the Pre-Post-Then design, the second to the Pre-Post&then design, the third to the Post-Then design, and the fourth to the Post&then design. The process continued until all participants were assigned to a design. Subsequent assignments picked up where the prior class roster left off.

Table 1. Evaluation Design by Group

Group	Evaluation Design					
Pre-Post-Then	Ob	R	Pr	XI	Po[a]	Th[b]
Pre-Post&then	Ob	R	Pr	XI	Po&Th[c]	
Post-Then	Ob	R	–	XI	Po[a]	Th[a]
Post&then	Ob	R	–	XI	Po&Th[c]	

Note. Ob = objective measure; Po = self-report posttest; Pr = self-report pretest; R = random assignment; Th = self-report thentest; XI = program.
[a] Separate posttest.
[b] Separate thentest.
[c] Combined posttest and thentest, where survey instructions asked participants to provide posttest and thentest responses before moving to next survey item.

Repeated measures of traditional self-report data were collected according to the particulars of the retrospective pretest design. In addition, data were collected pertinent to supporting individual research questions (e.g., objective performance measure and self-report control measure).

In support of the first research question, an objective pretest measure was incorporated across each of the four designs to validate participants' responses to then items, following Umble, Upshaw, Orton, and Matthews (2000) and Pratt, McGuigan, and Katzev (2000). Scores from a skill-based test that participants took approximately 1 week prior to the program provided the objective pretest measure. We related skill-based pre ratings to then ratings to establish the level of criterion validity of then ratings across the four designs. Given that the skill-based test had a published nomological net inclusive of the self-report measures, we compared the resulting validity data across the four designs to published results. The nomological net served as a benchmark and assisted us in assessing which designs produced the most criterion valid then ratings.

In support of the second research question, a control condition was included to identify to what extent participants reported improvement in a construct outside the scope of the program, following Taylor et al. (2009) and Sprangers and Hoogstraten (1989). Similar to the internal referencing strategy described in Haccoun and Hamtiaux (1994), the control measure assessed individuals on a measure that had discriminant validity with the construct designed to change as a result of participation in the leadership development program. This technique allowed us to quantify the level of inflationary bias across designs and assess which designs produced the most biased measures of program effectiveness.

In support of the third research question, which considered whether the self-report pre ratings were subject to experience limitation, we used the objective pretest scores collected in support of the first research question to conduct validity analyses on the self-report pre ratings. We compared the validity data of the self-report pre ratings against the same benchmark used to validate the self-report then ratings.

Program

The leadership program evaluated focused on the Situational Leadership® II (SLII) model (Blanchard, Zigarmi, & Zigarmi, 1985). The program trained managers how to identify the needs of their employees and tailor their leadership style to different situations. During the program, managers also reviewed how they typically responded to employee needs and the effectiveness of their leadership responses. We expected that through the experience of training, participants would reframe how they judged their leadership competence.

Participants

Participants from 15 classes of the same SLII training program provided the data for the study. Certified trainers used by the same international training company facilitated the classes. Across the 15 classes, there were 163 participants. Approximately 15% of the participants returned surveys with excessive missing data, leaving an effective sample size of 139 and an approximate response rate of 85%.

Measures

Participants completed scales from two leadership competency instruments: (a) the Survey of Management Practices Self (SMP) and (b) The Leader Behavior Analysis II® Self (LBAII). Scores from the SMP provided self-report data on participants' perceived leadership competence and served as the study's self-report measures. Scores from the LBAII provided an assessment of leadership competence based on the underlying model of the program (SLII) and served as the study's objective performance measure.

Self-report measures. The SMP consists of 145 items designed to measure perceived managerial competence on 11 skills and 12 attributes (Wilson, 2006). Wilson (1978) provided evidence indicating that SMP scale scores demonstrated dimensionality, item accountability, reliability, reproducibility, as well as the ability to discriminate between managers. The instrument incorporates a 7-point Likert-type scale with appropriate anchors for a competency-based scale (Shipper, 1995), with 1 indicating *never or to a very small extent*, 2 indicating *almost never or to a little extent*, 3 indicating *sometimes or to a less than average extent*, 4 indicating *average*, 5 indicating *often or to a more than average extent*, 6 indicating *almost always or to a large extent*, and 7 indicating *always or to a very great extent*. Responses to SMP items represent respondents' subjective perceptions of their competencies and their workplace.

Participants completed the following SMP scales: Clarification of goals and objectives (SMPa, 7 items), Upward communication (SMPb, 8 items), and Work involvement (SMPp, 5 items). Self-responses to SMPa, SMPb, and SMPp scales have yielded reliability coefficients of .87, .78, and .90, respectively (Wilson, 1978). Responses to SMPa items (e.g., "Tells group members how their jobs, work, and goals relate to organization's goals") and SMPb items (e.g., "Encourages people to express their opinions and participate in decisions") served as the study's self-report program-related measures because goal clarification and upward communication were related to material covered in the training and were therefore expected to change as a result of participation in the program. Responses to SMPp items (e.g., "The work is stimulating") served as the study's self-report control measure because the training program content did not include work absorption and therefore we did not expect participants to change their perceptions of work absorption as a result of participation in the program.

Objective performance measure. The Leader Behavior Analysis II Self (LBAII) is a 20-item instrument based on the SLII model. The LBAII contains five different leadership scenarios for each of the four development levels (i.e., D1—low competence/high commitment, D2—low commitment/some competence, D3—variable commitment/high competence, and D4—high commitment/high competence) represented in the model. Following each scenario, survey questions task respondents to select the most appropriate leader response from four possible choices. Each of the possible responses represents one of the four leadership styles (i.e., directing, coaching, supporting, and delegating) from the SLII model.

Although the LBAII yields six measures, Zigarmi, Edeburn, and Blanchard (1997) considered the Effectiveness Score (LBAIIe) to be the most important measure because it is the "raison d'etre for

the model'' and is correlated to ''key managerial behaviors researched by other authors of management'' (p. 28). Effectiveness scores range from 20 to 80, with each item having a maximum score of 4 (the most appropriate theoretical leadership response to the development level presented) and a minimum score of 1 (the poorest leadership response based on the SLII model). For the current study, LBAIIe scores served as the objective measure of program effectiveness because they provided a criterion-based indicator of respondents' ability to correctly identify the most appropriate leadership response according to the development level of a follower on a particular task, according to the SLII model.

Punch (as cited in Zigarmi, Edeburn, & Blanchard, 1997) examined reliability of LBAIIe scores. Conducting a Rasch analysis, he found that 15 of the 20 items fit the response model very well, 2 items overdiscriminated, and 3 items underdiscriminated. McDermot (as cited in Zigarmi et al., 1997) found that LBAIIe scores discriminated between managers who attended a 3-day Situational Leadership training workshop and a matched group of managers who did not. Zigarmi et al. (1997) established nomological validity of the LBAII by relating LBAIIe scores to SMP scores based on a sample of 552 subordinates who assessed their managers using both the LBAII and SMP. Relevant to this study are the validity data for the SMPa, SMPb, and SMPp. As evidence of convergent validity, correlations between LBAIIe and SMP scores were .389 for the SMPa and .388 for the SMPb. As evidence of discriminant validity, Zigarmi et al. reported no statistical difference in SMPp scores between individuals who scored above or below the normal range (50–58) of the LBAIIe.

Procedures

Within each group, participants completed the LBAII and select scales of the SMP (i.e., SMPa, SMPb, and SMPp). Participants in all four groups completed the LBAII approximately 1 week prior to attending training. During training, participants completed SMP tests according to their retrospective pretest design group assignment. The Pre-Post-Then group completed a pretest at the beginning of training, a posttest as training was completing, and a thentest immediately following the posttest. Pre-Post&then group completed a pretest at the beginning of training and a combined posttest/thentest as training was completing. The Post-Then completed a posttest as training was completing and a thentest immediately following the posttest. The Post&then group completed a combined posttest/thentest as training was completing. In the Pre-Post-Then and Post-Then group, participants received the posttest and thentest in separate envelopes that indicated the prescribed order the tests were to be completed. Survey instructions directed participants to complete each and place it in its respective envelope before proceeding to the next assessment. The time period between pre and post testing was approximately 4 days.

Across all groups, thentest item directions followed guidelines established by Howard et al. (1979) and asked participants to reassess their pre-program abilities. For participants in the Pre-Post-Then and Pre-Post&then groups, who had completed a pretest, thentest item directions included additional guidelines established by Mezoff (1981) and asked participants not to recall their prior answers or worry whether their reevaluated ratings agreed or disagreed with prior ratings.

The layout of the stand-alone pretest, posttest, and thentest included two columns. The first column contained the survey items. The second column contained the response options. The layout of the combined posttest/thentest included three columns, as in Lamb and Tschillard (2005). The first column contained the survey item; the second column contained posttest response options; and the third column contained thentest response options. In the combined posttest/thentest, survey instructions asked participants to respond to each survey item twice before moving on to the next survey item, as in Howard et al. (1979). First, they were to report how they perceived themselves to be at present (Post). Next, they were to report how they perceived themselves to have been as they were commencing the leadership program (Then).

Data Analysis

To answer our first research question, we examined the criterion validity of the study's thentest measures and compared results between groups and to established benchmarks. We correlated then ratings from the study's self-report program-related measures (SMPa and SMPb) to pre ratings from the study's objective performance measure (LBAIIe). We transformed resulting correlations and related correlations reported in Zigarmi et al. (1997) into validity estimates (r') using Cohen's (1988) formula:

$$r' = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right).$$

Statistical differences between the validity estimates were calculated using the z statistic (Hinkle, Wiersma, & Jurs, 2003):

$$z = \frac{r'_1 - r'_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}},$$

where n equals group size. We computed practical differences as a simple function of the difference in validity estimates, defined by Cohen (1988) as q.

We examined discriminant validity of then ratings from the study's self-report control measure (SMPp) by replicating analyses conducted by Zigarmi et al. (1997), who categorized SMPp scores by LBAIIe (i.e., high, norm, and low) and compared the low and high groups to determine the relationship between LBAIIe and SMPp scores. We compared SMPp then ratings between individuals who scored above 58 on the LBAIIe pretest to those who scored below 50.

To answer our second research question, we evaluated statistical and practical measures of program effectiveness on each of the study's self-report measures (SMPa, SMPb, and SMPp) across designs. We used repeated measures t tests to test for statistical differences between post and then ratings. For each t test conducted, we computed a parallel effect size to examine the practical significance of differences between post and then ratings. We used the formula for computing d for repeated measures design, as advocated by Dunlap, Cortina, Vaslow, and Burke (1996):

$$d_c = t_c[2(1-r)/n]^{1/2},$$

where t_c is t for correlated measures, r is the correlation between measures, and n is the sample size per group.

To answer our third research question, we combined pre ratings from the Pre-Post-Then and Pre-Post&then groups for each of the study's self-report measures (SMPa, SMPb, and SMPp) and assessed the criterion validity of pre ratings. Conducting analyses similar to those described for Research Question 1, we assessed the criterion validity of SMP pre ratings and compared those results to the criterion validity of SMP then ratings and to benchmarks derived from the study by Zigarmi et al. (1997). We also examined the effect of response–shift in the Pre-Post-Then and Pre-Post&then groups. Conducting analyses similar to Research Question 2, we estimated the effect of response–shift bias by conducting repeated measures t tests on then and pre ratings and computing d_c.

Findings

This study examined the impact of retrospective pretest design on several measurement outcomes. First, the study examined the criterion validity of then ratings and compared the results between

Table 2. Thentest Validity Estimates (*r*') and Differences (*q*) for Self-Report Program-Related Measures (SMPa and SMPp)

Evaluation Design	SMPa	SMPb	$(\lvert\bar{q}\rvert)$
Pre-Post-Then	.425	.405	.004
	−.004	.005	
Pre-Post&then	.141	−.188*	.439
	.280	.598	
Post-Then	.297	.290	.122
	.124	.120	
Post&then	.003*	.303	.262
	.418	.107	
Zigarmi et al. (1997)	.421	.409	

Note. SMPa = SMP Clarification of goals and objectives. SMPb = SMP Upward communication; *q* = difference in validity estimate as compared to Zigarmi et al.
* $p < .05$.

groups and to published validity data. Second, the study evaluated measures of practical significance derived from post-program survey data across each of the study's self-report measures. Third, the study analyzed self-report pre ratings to examine the effect of experience limitation.

Then Ratings

Table 2 outlines the validity estimates for then ratings from the self-report program-related measures (SMPa and SMPb). Across retrospective pretest design groups, validity estimates ranged from .003 (Post&then) to .425 (Pre-Post-Then) for the SMPa and from −.188 (Pre-Post&then) to .405 (Pre-Post-Then) for the SMPb. These differences were statistically (ps < .05) and practically significant (qs > .40). In addition, the SMPb validity estimate for the Pre-Post&then group (-.188) was statistically and practically different (ps < .05, qs < .40) than the SMPb validity estimate for the Post-Then (.290) and the Post&then group (.303).

With the post-program assessment formatted as two separate surveys, the validity estimates of the SMPa and SMPb then ratings were comparable to the validity data derived from the study by Zigarmi et al. (1997). The Pre-Post-Then design yielded the most comparable thentest validity data ($|\bar{q}| = .004$), followed by the Post-Then design ($|\bar{q}| = .122$). With the post-program self-report survey formatted on a single survey, at least one of the two SMP measures failed to compare to the validity data derived from Zigarmi et al. The Pre-Post&then design yielded the least comparable thentest validity data ($|\bar{q}| = .439$), followed by the Post&then design ($|\bar{q}| = .262$).

Irrespective of pretest administration, designs that incorporated separate post-program surveys produced the most comparable levels of criterion validity for program-related thentest data. However, when the post-program assessment was formatted in two separate surveys, the design that incorporated a traditional pretest (Pre-Post-Then) yielded slightly more comparable thentest validity data than the design that did not (Post-Then). Conversely, when the post-program assessment was formatted in a single survey, the design that incorporated a traditional pretest (Pre-Post&then) yielded slightly less comparable thentest validity data than the design that did not (Post&then). Figure 1 depicts a slight interaction between self-report pretesting and post-program survey format on validity coefficient differences across the self-report program-related measures.

Across retrospective pretest designs (see Table 3), the self-report control measure (SMPp) did not relate to performance in LBAII scores, when considering statistical significance (ps = .156–.855). However, when considering measures of practical significance, performance on LBAII resulted in effect sizes ranging from nonexistent (.002) to medium (.124). A medium effect (.121) was detected

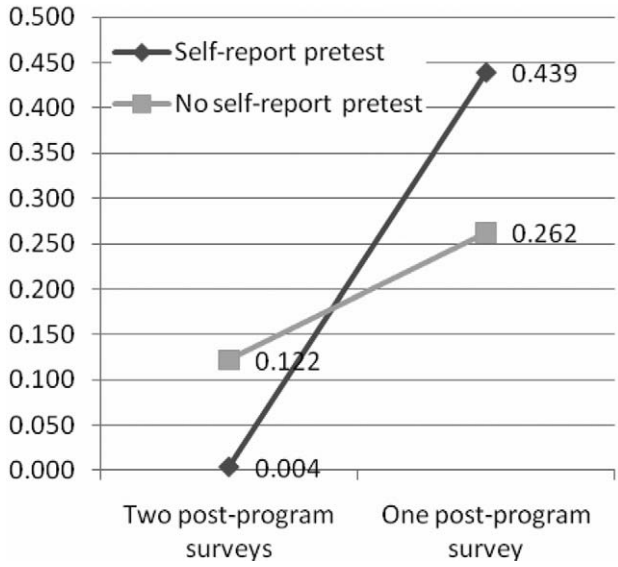

Figure 1. Average validity coefficient bias ($|\bar{q}|$) across self-report program-related measures (SMPa and SMPb) as a function of post-program survey format and self-report pretesting. SMPa = SMP Clarification of goals and objectives; SMPb = SMP Upward communication.

Table 3. Analysis of Variance Between Then Ratings From Self-Report Control Measure (SMPp) by Retrospective Pretest Design

Group	*F*	df_1	df_2	*p*	η^2	Low	High
Pre-Post-Then	.034	1	22	.855	.002	6.086	6.160
Pre-Post&then	2.211	1	16	.156	.121	4.440	5.585
Post-Then	.682	1	19	.419	.034	5.720	6.125
Post&then	.642	1	19	.433	.033	5.333	5.866

Note. SMPp = SMP work involvement.

in the Pre-Post&then design, small effects were detected in the Post-Then and Post&then designs (.034 and .033, respectively), and virtually no effect was detected in the Pre-Post-Then (.002) group. Given the findings of the study by Zigarmi et al. (1997), SMPp then ratings were most valid in the Pre-Post-Then group and least valid in the Pre-Post&then group.

Measures of Program Effectiveness

Table 4 presents means, standard deviations, and training effect sizes for each of the self-report measures. Measures of program effectiveness on the control measure (SMPp) were different across retrospective pretest designs, as measured by statistical (*p*) and practical (*d*) differences between post and then ratings. Values of *p* ranged from nonsignificant (.08) to significant (<.01) and values of *d* ranged from small (.23) to moderate (.42). Although only the Pre-Post-Then design showed no statistically significant program effect ($p = .08$, $d_c = .23$) on the control measure, the Post-Then yielded the next smallest effect size on the control measure ($d_c = .34$). The Pre-Post&then and Post&then designs yielded the largest effect sizes (d_cs = .42 and .41, respectively).

Table 4. Descriptive Statistics for Self-Report Program-Related (SMPa and SMPb) and Control (SMPp) Measures and Training Effects

		Pretest *M* (Pretest *SD*)			Posttest *M* (Posttest *SD*)			Thentest *M* (Thentest *SD*)			Posttest–Thentest *t* (Training Effect Size)		
Design	*n*	SMPa	SMPb	SMPp	SMPa	SMPb	SMPp	SMPa	SMPb	SMPp	SMPa	SMPb	SMPp
Pre-Post-Then	38	5.40 (.84)	5.70 (.79)	6.12 (.93)	5.90 (.71)	6.17 (.62)	6.27 (.84)	5.21 (1.05)	5.43 (1.14)	6.02 (.95)	5.61*** (.71)	5.55*** (.67)	1.75 (.23)
Pre-Post&then	37	5.51 (.96)	5.65 (.70)	6.00 (.94)	6.24 (.48)	6.12 (.58)	6.24 (.80)	5.30 (.77)	5.41 (.75)	5.74 (1.34)	8.10*** (1.39)	5.75*** (1.05)	3.42** (.42)
Post-Then	31				6.13 (.71)	6.15 (.70)	6.25 (.89)	5.44 (.98)	5.58 (.88)	5.84 (1.22)	4.80*** (.77)	4.37*** (.71)	3.44** (.34)
Post&then	33				5.93 (.80)	6.23 (.61)	6.21 (.90)	5.24 (.91)	5.47 (.67)	5.81 (1.05)	5.68*** (.80)	8.10*** (1.23)	2.76* (.41)

Note. SMPa = SMP Clarification of goals and objectives; SMPb = SMP Upward communication; SMPp = SMP Work involvement. Standard deviations and effect sizes in parentheses. All *t*'s are from repeated measures *t* tests. Effect Size = *d* for repeated measures design (d_c).

*$p < .05$.

**$p < .01$.

***$p < .001$.

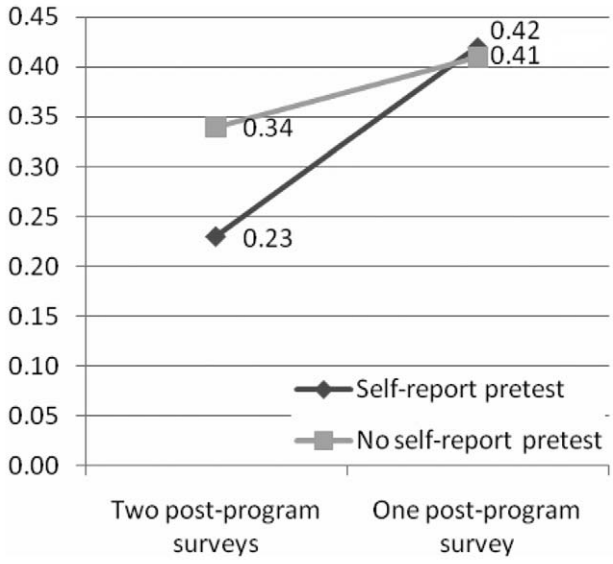

Figure 2. Program effect (d_c) on self-report control measure (SMPp) as a function of post-program survey format and self-report pretesting. SMPp = SMP Work involvement.

With the post-program assessment formatted as a single survey, the related designs yielded *larger* effect sizes on the control measure than designs that incorporated separate post-program surveys. This finding was consistent irrespective of pretest administration. However, with the post-program assessment formatted as two separate surveys, the design that incorporated a traditional pretest (Pre-Post-Then) yielded slightly *smaller* effect sizes on the control measure than the design that did not (Post-Then). Figure 2 demonstrates the relationship between self-report pretesting and post-program survey format on training effect sizes derived from the control measure.

Statistically significant differences between post and then ratings were found for the remainder of the self-report measures (SMPa and SMPb), across all retrospective pretest designs, as expected. However, across retrospective pretest designs, there were practical differences in the training effect sizes across the program-related measures. The Pre-Post-Then design yielded the most conservative effect sizes across program-related measures ($\overline{d}_c = .69$), followed by the Post-Then design ($\overline{d}_c = .74$). The Pre-Post&then design yielded the least conservative effect sizes across program-related measures ($\overline{d}_c = 1.22$), followed by the Post&then design ($\overline{d}_c = 1.05$).

With the post-program survey formatted as two separate surveys, the related designs yielded more conservative effect sizes across the program-related measures than designs that incorporated a single post-program survey. This finding was consistent irrespective of pretest administration. However, with the post-program assessment formatted as two separate surveys, the design that incorporated a traditional pretest (Pre-Post-Then) yielded slightly *more* conservative effect sizes across the program-related measures than the design that did not (Post-Then). Conversely, when the post-program assessment was formatted in a single survey, the design that incorporated a traditional pretest (Pre-Post&then) yielded slightly *less* conservative effect sizes across the program-related measures than the design that did not (Post&then). Figure 3 depicts a slight interaction between self-report pretesting and post-program survey format on effect sizes derived from the self-report program-related measures.

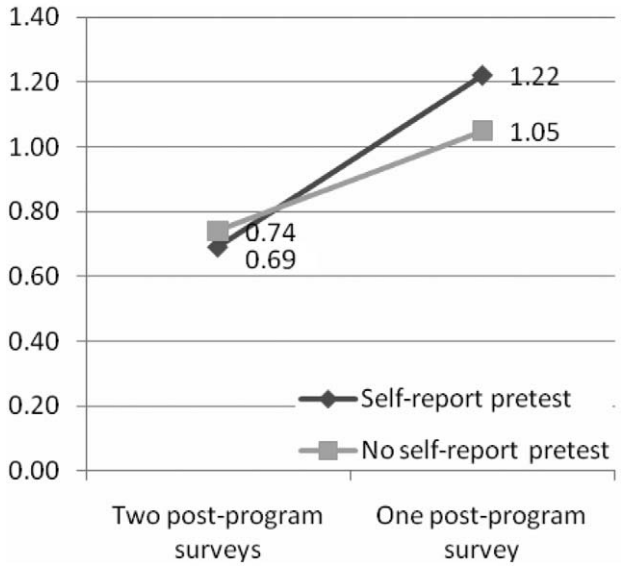

Figure 3. Average program effect $(\bar{d}_c)$ across self-report program-related measures (SMPa and SMPb) as a function of post-program survey format and self-report pretesting. SMPa = SMP Clarification of goals and objectives; SMPb = SMP Upward communication.

Pre Ratings

Validity estimates for SMPa pre ratings was .014 and .004 for SMPb pre ratings. These validity estimates were statistically and practically different than the validity coefficients for the SMP then ratings from the Pre-Post-Then group (ps < .05, qs = .40). These validity estimates were also statistically and practically different from the validity estimates derived from the study by Zigarmi et al. (1997). The average difference in validity estimates $(|\bar{q}|)$ was large (.403), reflecting a lack of expected criterion validity between participants' self-report and objective performance program-related pre ratings. In contrast, the validity estimates for pre ratings from the self-report control measure were generally consistent between groups and with the findings from the study by Zigarmi et al., where performance on the LBAII did not have a statistically significant relationship with SMPp scores.

As it relates to response–shift, the study found approximately one quarter of a standard deviation between then and pre ratings in both the Pre-Post-Then and Pre-Post&Then group across the majority of the study's measures. In the Pre-Post-Then group, differences were found in the SMPa ($t = 1.62$, $p = .11$, $d_c = .19$) and SMPb ($t = 2.39$, $p = .02$, $d_c = .24$) but not in the SMPp measure ($t = .48$, $p = .65$, $d_c = .06$). In the Pre-Post&Then group, differences were found in the SMPa ($t = 1.64$, $p = .11$, $d_c = .23$), SMPb ($t = 2.16$, $p = .04$, $d_c = .34$), and SMPp measure ($t = 1.94$, $p = .06$, $d_c = .22$).

Conclusions and Discussion

We conclude that as study participants were beginning training, they lacked sufficient information to accurately rate their pre-program program-related abilities. The near zero correlations found between program-related self-report and objective performance pretest measures provide support for this conclusion. This finding is also in line with prior literature. Pratt et al. (2000) found near zero

correlations between subjective and objective pretest measures of mothers' knowledge of early childhood development. Nisbett and Wilson (1977) argued that individuals cannot reliably use *introspection* to gain accurate assessment of their higher order cognitive processes, such as abilities. Aiken and West (1990) identified experience limitation as a potential source of bias in self-report pre ratings.

It does appear, however, that individuals can use *retrospection* to gain accurate assessments of program-related abilities after participating in a program that informs them of their abilities. Our study found that, when participants assessed their pre-program abilities after receiving information as to the effectiveness of their leadership behaviors, the criterion validity of their responses was more comparable to established benchmarks than the criterion validity of those without such information. This finding is consistent with Pratt et al. (2000), who found that pre-program self-ratings provided after training were more consistent with objective ratings than ratings provided before training. As in Pratt et al., we attribute the positive differential between the criterion validity of then and pre ratings to response–shift (Howard, 1980). We conclude that participants experienced a shift in their internal standards as a consequence of participating in the leadership development program. In our study, the consequence of response–shift resulted in more criterion valid assessments of participants' pre-program program-related abilities.

It also appears that the degree to which the participants' retrospective assessments of their pre-program abilities conformed to established benchmarks was a function of self-report pretest administration (with or without a pretest) and post-program survey format (separate or combined surveys). Moreover, measures of program effectiveness derived from post-program self-report measures were a function of self-report pretesting and post-program survey format. This latter finding is especially important because ''effect sizes are often evaluated relative to the cost of a program, political and social values, and the availability of alternative treatments'' (May, 2004, p. 532).

This study found that designs that included the administration of a posttest separate from a thentest yielded thentest validity data that were most comparable to established benchmarks and the least biased program effects (both program-related and control). Conversely, designs that incorporated a single post-program survey, with adjacent post and then items, yielding thentest validity data that were least comparable to established benchmarks and the most biased program effects (both program-related and control). Our results confirm the theory informed by Schwarz (1996) that arranging post and then items side by side may facilitate implicit theories of change, self-presentation, or effort justification, thereby introducing bias in post-program measurement outcomes. We conclude that the effect of administering a standalone thentest allowed study participants to provide more valid ratings than their counterparts who responded to post and then items simultaneously on the same survey.

This study also found a slight interaction between the effects of self-report pretesting and post-program survey format. Self-report pretesting had a slight positive effect on post-program survey data resulting from the separate administration of post and then items, with data yielding thentest validity coefficients that were slightly more comparable to established benchmarks and slightly less biased program effects. Conversely, self-report pretesting had a slight negative effect on post-program survey data resulting from the administration of a single survey containing adjacent post and then items, with data yielding thentest validity coefficients that were slightly less comparable to established benchmarks and slightly more biased program effects. The latter finding is consistent with Terborg and Davis (1982), who found that gains between post and then responses to a single post-program survey were higher for participants who also took a traditional pretest. Although the findings of the current study are not sufficient to provide definitive statements, the positive self-report pretest effect on data from Pre-Post-Then design could be tentatively associated with a sensitization effect and the negative self-report pretest effect on data from Pre-Post&then design could be tentatively associated with a compliance with implicit demands, as in Sprangers and Hoogstraten (1989).

Limitations and Recommendations for Future Research

This study focused on the self-perceptions of SLII participants, as measured by select scales from the Survey of Management Practices. Generalizability of these findings is therefore limited to similar measures from like populations. Future research should replicate this study to determine whether these findings are reproducible in other populations and with other survey measures. We recommend that future studies incorporate empirically based validity benchmarks and use survey measures with established psychometric properties. Such studies might also be designed as mixed-methods and solicit written comments, as in Howard et al. (1979), or include cognitive interviews with a small group of participants (see Willis, 2005) to shed light on the mechanisms underlying post-program response differences.

In the current study, participants completed paper forms of the assessments. Therefore, there is no guarantee that participants in the Post&then and Pre-Post&then group adhered to the post-training assessment directions that instructed them to respond to each survey item in a prescribed order (then after post) before moving on to the next survey item. Similarly, it is possible that participants in the Post-Then and Pre-Post-Then group reviewed their post responses as they were providing their then ratings, despite directions that instructed them to place each test back in its respective envelope before going on to the next. To control for these limitations, future research should replicate the current study using an online survey format.

Although we recognize that not all training evaluation facilities have computer workstations, the administration of computer-based assessments from handheld devices or computer workstations could facilitate enforcement of survey instructions. Such control would make it impossible for participants who complete separate posttests and thentests to review posttest responses as they were completing thentest items or for participants who complete combined posttests and thentests to respond to items outside of the prescribed order. Not only would such a study provide greater control of the test administration, it would provide the opportunity to determine whether the results found in the current study are generalizable to an online administration. Other possible studies of survey manipulation (either electronic or paper) include the order effect of post-program test administration (i.e., posttest before thentest vs. thentest before posttest) as in Terborg and Davis (1982) and the effect of responding to two separate post-program surveys versus a single survey with post and then questions on separate ''pages'' of a single survey. Such studies should also consider controls to enforce desired survey procedures (e.g., controlling participant access to post ratings as they complete then ratings).

Limitations in group sizes prohibited examination of the internal structure of the post-program survey data. This study makes no claim as to the factor analytic integrity of the data across measurement occasion and retrospective pretest design group. Future research should test for measurement invariance of data resulting from retrospective pretest designs. It would be interesting to note whether the same measurement variance issues between pre and post ratings (Pitts, West, & Tein, 1996) are evident between then and post ratings.

Given the constraints of the leadership program evaluated, we did not assess the validity of post ratings. The validity claims identified in this study are therefore limited to measures of program effectiveness and then ratings. Future research should determine whether self-report pretesting and post-program survey format bias the validity of post ratings.

In the current study, participants took a performance test prior to training, the resulting scores of which provided evidence of criterion validity of then ratings. As such, generalizability of the findings from the current study may be limited as findings may be confounded by the assessment of the performance measure. Empirical evidence identifying the effect of taking a performance test on post-program survey ratings was found in Hoogstraten (1985) and Howard, Schmeck, and Bray (1979). However, in both studies, participants completed the performance test as training was

commencing, whereas in the current study, participants took the performance test approximately 1 week prior to training. Future research should examine techniques to evaluate criterion validity of self-report measures to determine which methods produce the least amount of bias and interference with the program.

The current study did not consider other retrospective self-reporting methods that do not involve a thentest. Lam and Bengo (2003) compared such methods to the practice of collecting post and then ratings from a single post-program survey. Future research could consider replicating a variant of the study by Lam and Bengo whereby participants assigned to the retrospective pretest group could complete separate posttest and thentest surveys, with survey procedures inhibiting them from referring to posttests as they complete thentests.

Future research involving the retrospective pretest should describe the format and administration of pre- and post-program measurement tools. With the exception of teaching papers that illustrate single post-program surveys (e.g., Lamb & Tschillard, 2005; Raidl et al., 2004), contemporary descriptions of post-program survey techniques are limited. Across articles in the *American Journal of Evaluation*, article of Hill and Betz (2005) is unique in that it explicitly defined the procedure for collecting post and then ratings. Findings from this study indicate that researchers using the thentest should explicitly report how pre- and post-program survey data are collected and consider how such techniques may influence resulting measures of program effectiveness based on self-report measures.

Recommendations for Program Evaluation

Given the findings of our study, we offer three recommendations for program evaluators using retrospective pretest designs. First, administer post and then items as separate surveys with appropriate survey administration procedures. Second, administer a self-report pretest when possible. Third, incorporate evaluation techniques to determine the validity of resultant data.

Separate Posttest and Thentest Surveys

We recognize the convenience of combining post and then response options on a single post-program survey. However, with that convenience come limitations. This study found that the validity of then ratings resulting from a single post-program survey failed to consistently conform to established benchmarks. When considering the suggestion of Hill and Betz (2005) that post and then ratings collected from a single post-program instrument can be useful if the relative ranking of participants' change scores are of interest, we question the validity of analyzing derived change scores that may be based on observed scores with no convergent validity with the program being evaluated. We add a caveat to the suggestion by Hill and Betz and assert that such analyses can be considered only if the validity of the data collected through the post-program survey can be ascertained. In our study, such ranking would not have been plausible for at least one of the study's program-related measures in the Pre-Post&then and Post&then groups.

When the goal of an evaluation incorporating a thentest is to ascertain whether a program was effective or not, we recommend that post-program survey procedures include the administration of a posttest separate from a thentest and that survey procedures inhibit participants from being able to refer to the posttest while completing the thentest, as in Terborg and Davis (1982). In addition to survey directions, evaluators should consider computer controls, physical manipulation of the evaluation environment (i.e., thentests administered after posttests turned in to the evaluator), or administering the thentest at a different time from the posttest to guard against participants referring to a posttest as they complete a thentest. Administering the posttest at a different time from the thentest has the added advantage of minimizing "respondents' ability to heuristically provide lower retrospective pretest rating than posttest ratings" (Taylor et al., 2009, p. 42).

When the goal of an evaluation incorporating a thentest is to provide participants an opportunity to reflect on how they may have changed as a function of participating in a program by way of post and then ratings (Hill & Betz, 2005), we recommend administering post and then items via two separate surveys. Findings from this study indicate that a single post-program survey may encourage participants to manipulate their reassessment of their pre-program perceptions to show an improvement in KSAs even if no such change occurs. Such manipulation could give participants a biased view of program change and their pre-program abilities.

Self-Report Pretesting

We recognize that the administration of a self-report pretest may not be plausible in all evaluation scenarios. As noted in Hill and Betz (2005), evaluators may feel that asking participants to reveal information about their abilities before a program has begun may be unrealistic or offensive until the establishment of rapport, especially if the information requested is sensitive. Additionally, Klatt and Taylor-Powell (2005) noted resource constraints and late arrivers as reasons why the administration of traditional pretest may not always be plausible. There is also the possibility that self-report pretesting may introduce carryover, interference, and sensitization effects in data collected post-program (cf., Lam & Bengo, 2003).

We recommend that evaluators administering thentests consider the potential issues associated with self-report pretesting and administer self-report pretests when possible. In the current study, there appeared to be no negative effect to self-report pretesting when participants completed separate posttests and thentests. Administering a thentest and pretest also facilitates the detection of response–shift, which allows for greater understanding of reactions to the program (Sprangers & Hoogstraten, 1989; Terborg & Davis, 1982). In addition, a Pre-Post-Then design is a simple extension of the traditional pre-post design, which may provide evaluators the legitimacy they seek when interfacing with program funders or colleagues who are accustomed to more traditional forms of research (cf., Betz & Hill, 2006).

Evidence of Validity

We recommend that evaluators administering thentests collect additional data to validate responses to self-report surveys. When assessing program effectiveness, it is a well-established principle to use multiple methods. When considering the need to establish the validity of outcome measures with the importance of incorporating multiple methods to access change, one sees that the issues are reciprocal. By incorporating concurrent methods to assess program effectiveness, an evaluator can build validity evidence for self-report measures by identifying how well they converge or discriminate with objective or other report measures. However, such validity evidence is limited if it cannot be placed within the extant of existing research, either theoretical or empirical. By comparing criterion relationships to published research, evaluators can confirm that the meaning of their self-report data is consistent with what is known in the literature. Without such evidence of validity, evaluators using a retrospective pretest design may presume, for example, that then ratings are more accurate if they result in larger program effects than those based on traditional pre ratings. However, we agree with Hill and Betz (2005) that such presumptions are unfounded.

A sound design incorporating the thentest should incorporate multiple measures, supported by an established nomological net, to adequately assess program effectiveness. In addition, evaluators should follow sound survey design principles and systematically consider potentials source of bias and devise strategies to counter related threats to validity. Although such a design goal is a consideration for any evaluation, it is a necessity for studies incorporating a thentest lest they fall victim to inherent concerns related to retrospective accounts.

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

References

Aiken, L. S., & West, S. G. (1990). Invalidity of true experiments: Self-report pretest biases. *Evaluation Review*, *14*, 374-390.

Betz, D. L., & Hill, L. G. (2006). Real world evaluation. *Journal of Extension*, *44*. Retrieved November 20, 2009, from http://www.joe.org/joe/2006april/rb9.php

Blanchard, K. H., Zigarmi, D., & Zigarmi, P. (1985). *Leadership and the one minute manager.* New York, NY: Morrow Publishers.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago, IL: RandMcNally.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Furby, L. (1970). How should we measure ''change'' or should we? *Psychological Bulletin*, *74*, 68-80.

Deutsch, M., & Collins, M. E. (1951). *Interracial housing: A psychological evaluation of a social experiment.* Minneapolis, MN: University of Minnesota Press.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures. *Psychological Methods*, *1*, 170-177.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (pp. 41-48). New York, NY: Academic Press.

Haccoun, R. R., & Hamtiaux, T. (1994). Optimizing knowledge tests for inferring learning acquisition levels in single group training evaluation designs: The internal referencing strategy. *Personnel Psychology*, *47*, 593-604.

Hill, L. G., & Betz, D. L. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation*, *26*, 501-507.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Hoogstraten, J. (1985). Influence of objective measures on self-reports in a retrospective pretest-posttest design. *Journal of Experimental Education*, *53*, 207-210.

Howard, G. S. (1980). Response-shift bias: A problem in evaluating programs with pre/post self-reports. *Evaluation Review*, *4*, 93-106.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal validity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement*, *3*, 1-23.

Howard, G. S., Schmeck, R. R., & Bray, J. H. (1979). Internal validity in studies employing self-report instruments: A suggested remedy. *Journal of Educational Measurement*, *16*, 129-135.

Klatt, J., & Taylor-Powell, E. (2005, October). *Synthesis of literature relative to the retrospective pretest design*. Paper presented at the 2005 Joint CES/AEA Conference, Toronto, Canada.

Lam, T. C. M., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation*, *24*, 65-80.

Lamb, T. A., & Tschillard, R. (2005, Spring). Evaluating learning in professional development workshops: Using the retrospective pretest. *Journal of Research in Professional Learning*. Retrieved October 19, 20005, from http://www.nsdc.org/library/publications/research/lamb.pdf

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and post testing periods. *Review of Educational Research*, *47*, 121-150.

May, H. (2004). Making statistics more meaningful for policy research and program evaluation. *American Journal of Evaluation*, *25*, 525-540.

Mezoff, B. (1981). How to get accurate self-reports of training outcomes. *Training and Development Journal*, *35*, 56-61.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-259.

Pearson, R. W., Ross, M., & Dawes, R. M. (1992). Personal recall and the limits of retrospective questions in surveys. In J. M. Tanur (Ed.), *Questions and survey questions, Meaning, memory, expression, and social interactions in survey* (pp. 65-94). New York, NY: Russell Sage.

Pelfrey, W. V., & Sr., & Pelfrey, W. V. Jr. (2009). Curriculum evaluation and revision in a nascent field: The utility of the retrospective pretest-posttest model in a homeland security program of study. *Evaluation Review*, *1*, 54-82.

Pitts, S. C., West, S. G., & Tein, J. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning*, *19*, 333-350.

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. *American Journal of Evaluation*, *21*, 341-349.

Raidl, M., Johnson, S., Gardiner, K., Denham, M., Spain, K., & Lanting, R. (2004). Use retrospective surveys to obtain complete data sets and measure impact in extension programs. *Journal of Extension*, *42*. Retrieved October 19, 2005, from http://www.joe.org/joe/2004april/rb2.shtml

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review*, *96*, 341-357.

Schwarz, N. (1996). *Cognition and communication, Judgmental biases, research methods, and the logic of conversation.* Mahwah, NJ: Lawrence Erlbaum.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93-105.

Sears, R. R., Maccoby, E. E., & Levin, H. (1957). *Patterns of child rearing.* Evanston, IL: Row Peterson.

Shipper, F. (1995). A study of the psychometric properties of the managerial skill scales of the survey of management practices. *Educational and Psychological Measurement*, *55*, 468-479.

Sprangers, M., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology*, *74*, 265-272.

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Gilding the outcome by tarnishing the past: Inflationary biases in retrospective pretest. *American Journal of Evaluation*, *30*, 31-34.

Terborg, J. R., & Davis, G. A. (1982). Evaluation of a new method for assessing change to planned job redesign as applied to Hackman and Oldham's job characteristic model. *Organizational Behavior and Human Performance*, *29*, 112-128.

Thorndike, R. L. (1949). *Personnel selection.* New York, NY: John Wiley.

Umble, K., Upshaw, V., Orton, S., & Matthews, K. (2000, June). *Using the post-then method to assess learner change.* Presentation at the AAHE Assessment Conference, Charlotte, NC, USA.

Walk, R. D. (1956). Self-ratings of fear in a fear-invoking situation. *Journal of Abnormal and Social Psychology*, *52*, 171-178.

Willis, G. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage.

Wilson, C. L. (1978). The Wilson multi-level management surveys: Refinement and replication of the scales. *JSAS: Catalog of Selected Documents in Psychology*, 8, 1707. Washington, DC: American Psychology Association.

Wilson, C. L. (2006). *The Clark Wilson group surveys: Management practices.* Silver Spring, MD: The Clark Wilson Group .

Zigarmi, D., Edeburn, C., & Blanchard, K. (1997). *Getting to know the LBAII®, Research, validity and reliability of the self and other forms* (4th ed.). Escondido, CA: Blanchard Training and Development, Inc.